Lidil
Revue de linguistique et de didactique des langues

61 | 2020
Le mépris en discours

Quel corpus pour l'identification des compétences des apprenants de niveaux intermédiaire et avancé ? Les cas de la cohérence et de la cohésion

What Corpus for Identifying Intermediate and Advanced Learners Abilities? The Case of Coherence and Cohesion

Antonin Brunet

Édition électronique
URL : http://journals.openedition.org/lidil/7424
DOI : 10.4000/lidil.7424
ISSN : 1960-6052

Éditeur
UGA Éditions/Université Grenoble Alpes

Édition imprimée
ISBN : 978-2-37747-195-9
ISSN : 1146-6480

Référence électronique
Antonin Brunet, « Quel corpus pour l'identification des compétences des apprenants de niveaux intermédiaire et avancé ? Les cas de la cohérence et de la cohésion », *Lidil* [En ligne], 61 | 2020, mis en ligne le 02 mai 2020, consulté le 02 mai 2020. URL : http://journals.openedition.org/lidil/7424 ; DOI : https://doi.org/10.4000/lidil.7424

Quel corpus pour l'identification des compétences des apprenants de niveaux intermédiaire et avancé ? Les cas de la cohérence et de la cohésion

What Corpus for Identifying Intermediate and Advanced Learners Abilities? The Case of Coherence and Cohesion

Antonin Brunet

1. Introduction et objet de l'étude

1 Il n'est pas rare, dans le domaine de la recherche, que des corpus initialement élaborés avec un objectif spécifique puissent être utilisés comme supports d'analyse dans le cadre d'autres recherches que celle initialement prévue[1]. En revanche, pour ces cas de figure, il semble bon de rappeler que chaque corpus est motivé par rapport à une fonction spécifique. Ainsi, son exhaustivité et son objectivité potentielles sont discutables, du fait que la méthodologie d'élaboration, motivée par un objectif spécifique de recherche, entrainera une sélection / une production de données ciblées et orientées (Bilger, 2008). En d'autres termes, dépendamment de l'objectif de recherche à partir duquel le corpus aura été constitué, les données ne seront représentatives que d'une partie de l'usage de la langue et la question de la représentativité sera alors à soumettre (Sinclair, 1991 ; Gadet, 2008). De ce fait, toute réutilisation d'un corpus existant avec un nouvel objectif de recherche doit être minutieusement réfléchie et certains aménagements seront probablement nécessaires.

2 À l'inverse, lorsqu'il s'agit d'élaborer un nouveau corpus pour répondre à un objectif de recherche, le chercheur encourt d'autres risques potentiels comme celui de sa

subjectivité (Rabatel, 2013). Ainsi, soucieux de vérifier une hypothèse, il peut inconsciemment constituer un corpus où cette dernière se vérifierait nécessairement par manque d'objectivité. Lors de la constitution d'un nouveau corpus, tout l'enjeu est alors de trouver le bon équilibre entre un corpus qui devient un support d'analyses objectif présentant des observables et la démarche circulaire où ces observables sont faussés (Bilger & Cappeau, 2008). Les corpus d'apprenants et leur exploitation, quel qu'en soit l'objectif, ne dérogent pas à ces constats.

3 Dans cet article, notre intention, en tant que jeune chercheur, est d'apporter une réflexion épistémologique sur l'outil de recherche que constitue le corpus à l'aide d'une expérience concrète qui nous aidera à mieux identifier les pièges dans lesquels tout chercheur risque de tomber. Nous proposons ainsi d'essayer de confirmer ou d'infirmer l'une des hypothèses de travail de notre thèse de doctorat à l'aide de deux corpus d'apprenants déjà constitués.

4 Pour mener à bien cette étude, nous tâcherons de nous concentrer sur les moyens employés par les apprenants de français langue étrangère (ci-après FLE) de niveaux intermédiaire à avancé pour construire un discours cohésif et cohérent. Notre hypothèse de travail est la suivante : pour ces composantes (cohérence et cohésion), il existerait un décalage entre les exigences du *Cadre européen commun de référence pour les langues* [CECRL] (Conseil de l'Europe, 2001) pour l'obtention d'un niveau et les compétences réelles d'un apprenant certifié.

5 Grâce à une analyse qualitative de quelques exemples représentatifs de chaque corpus, nous montrerons que la confirmation ou l'infirmation de notre hypothèse de départ est loin d'être évidente et que le corpus en tant qu'objet d'étude pour l'apport d'une réponse ne trouve sa validité scientifique que dans une suite de choix que le chercheur effectue depuis l'élaboration de son cadre expérimental jusqu'au traitement des données récoltées.

2. La cohésion et la cohérence dans le CECRL

6 Avant de procéder à l'identification des stratégies employées par les apprenants pour construire un discours cohérent et cohésif, nous avons voulu faire un tour d'horizon des descriptions que propose le CECRL pour ces deux notions. Le CECRL (chapitre 5.2) décompose les compétences communicatives langagières en trois catégories distinctes : les compétences linguistiques, sociolinguistique[2] et pragmatiques. Ces compétences pragmatiques sont elles-mêmes subdivisées en trois autres compétences : la compétence discursive, la compétence fonctionnelle et la compétence de conception schématique. C'est plus particulièrement la compétence discursive qui nous intéresse dans ce contexte. Le CECRL la définit comme celle permettant à l'apprenant d'ordonner les phrases en séquences afin de produire des ensembles cohérents. Notons que le cadre précise :

> Une grande partie de l'enseignement de la langue maternelle est consacrée à l'acquisition des capacités discursives. Dans l'apprentissage d'une langue étrangère, il est probable que l'apprenant commencera par de brefs énoncés d'une phrase seulement en général. Aux niveaux supérieurs, le développement de la compétence discursive dont les composantes sont inventoriées dans cette section devient de plus en plus important.
> Les échelles suivantes viennent illustrer certains aspects de la compétence discursive :

– souplesse ;
– tours de parole ;
– développement thématique ;
– cohérence et cohésion. (CECRL, 2001, p. 96-98)

7 L'échelle en question pour les deux composantes qui nous intéressent est la suivante :

- B1 : Peut relier une série d'éléments courts, simples et distincts en un discours qui s'enchaine.
- B2 : Peut utiliser un nombre limité d'articulateurs pour relier ses énoncés bien qu'il puisse y avoir quelques « sauts » dans une longue intervention. Peut utiliser avec efficacité une grande variété de mots de liaison pour marquer clairement les relations entre les idées.
- C1 : Peut produire un texte clair, fluide et bien structuré, démontrant un usage contrôlé de moyens linguistiques de structuration et d'articulation.
- C2 : Peut créer un texte cohérent et cohésif en utilisant de manière complète et appropriée les structures organisationnelles adéquates et une grande variété d'articulateurs.

8 Cette échelle nous indique ce qu'il faudrait attendre d'un apprenant selon son niveau de langue. Notons premièrement que le CECRL utilise alternativement les mots *textes* et *énoncés* pour référer aux productions des apprenants. En réalité, il n'est pas question ici de référer à un type de production en particulier : cette échelle semble être proposée indépendamment du canal de production et serait la même qu'il s'agisse d'une production orale ou écrite. Une deuxième remarque est que le CECRL semble mettre l'accent sur une progression d'ordre syntaxique et lexicale, en insistant d'une part sur l'enrichissement du nombre de connecteurs et d'articulateurs utilisés et d'autre part sur la complexification des structures organisationnelles choisies. Cependant, les moyens linguistiques cités par le CECRL se limitent aux connecteurs et aux articulateurs. Les structures organisationnelles adéquates ne sont quant à elles pas précisées.

9 Bien que le CECRL n'ait pas vocation à lister de manière exhaustive les procédés de construction de ces deux composantes — si tant est que faire se peut — il convient néanmoins d'apporter quelques références qui sauront nous aider pour notre analyse à venir. On sait par exemple que l'emploi des connecteurs et des articulateurs n'est pas suffisant comme gage de cohérence d'une production (Charolles, 1978 ; Reinhart, 1980). Salles (2006) et Vigier (2012) rappellent que ces marques de cohésion, mal ou trop utilisées, peuvent même aller jusqu'à détruire la cohérence. Enfin, cette échelle ne précise pas non plus les conditions nécessaires à la cohérence d'un discours. Dans un souci de synthèse, il semble difficile de toutes les rappeler dans cet article, mais nous pouvons tout de même mentionner à ce sujet les quatre méta-règles de cohérence de Charolles (1978), clairement résumées dans Salles (2006), qui consistent en la répétition, la progression, la non-contradiction et la relation. Une autre condition qu'il nous semble important de mentionner concerne celle du respect d'un des trois schémas de progression thématique — linéaire, à thème constant ou à thèmes dérivés — autour desquels tout discours se construit (Daneš, 1974 ; Carter-Thomas, 1999). Mentionnons enfin le fait qu'il existe deux types de cohérence qui, selon Salles (2006), sont susceptibles d'avoir entrainé des divergences d'opinions sur le caractère nécessaire ou non de la cohésion pour construire de la cohérence parmi les chercheurs spécialistes de ce domaine. Ainsi, il existe une cohérence locale, sur de courts énoncés, qui peut parfaitement se passer de marques de cohésion, et une cohérence globale, sur des énoncés plus longs, qui peut bien plus difficilement se dispenser de marques cohésives

mais dont la nature admettra « une force cohésive plus ou moins lâche » (Cornish, 2003).

3. Présentation des corpus et exemples de données

10 Ce rapide tour d'horizon effectué, nous cherchons maintenant à vérifier notre hypothèse initiale à l'aide de deux corpus d'apprenants existants. Rappelons que nous pensons qu'il existe un décalage entre les prérequis du CECRL et les compétences réelles d'un apprenant certifié. Rappelons également que l'échelle de progression proposée par le CECRL ne postule pas de différence selon que le canal soit écrit ou oral. Ainsi, les deux corpus à partir desquels nous avons travaillé sont :

- Le corpus oral Emo-FLE[3], réalisé dans le cadre des travaux du laboratoire FoReLLIS (EA 3816) sur les émotions sous la direction de Valetopoulos. Les productions ont été récoltées auprès d'apprenants allophones (grecs, chypriotes et polonais) de niveaux B1+ à C2. Le protocole visait à montrer une image à chaque apprenant et ce dernier devait la décrire dans un premier temps puis exprimer les émotions qu'il ressentait en la regardant. Cette consigne était répétée pour quatre images, présentées successivement par l'interviewer à l'apprenant.
- Un corpus écrit[4] constitué également au laboratoire FoReLLIS (EA 3816). Il s'agit de productions écrites collectées au Centre FLE de l'université de Poitiers lors du test de placement adressé aux étudiants de niveaux C1/C2. Le test est limité dans le temps et comporte trois parties, amenant les apprenants à produire des textes de typologies différentes. Le premier porte sur la description d'une image, le second sur la narration d'un évènement et le troisième sur une argumentation.

11 Nous allons donc observer les moyens mis en œuvre par les apprenants pour chacun de ces corpus et voir à la fois si ceux-ci semblent être fidèles à l'échelle du CECRL et s'ils sont les mêmes d'un corpus à l'autre.

3.1. Quelques échantillons du corpus oral

12 Dans la suite de cette partie, nous nous proposons, dans un souci de synthèse, de présenter quelques exemples représentatifs des données recueillies lors de la constitution du corpus Emo-FLE. L'analyse qualitative de ces exemples sera développée *infra* dans une partie dédiée.

(1)
[AP]euh + la premier- je pense la premier c'est + c'est l'été avec le soleil + c'est un endroit très calme on peut dire c'est Hawaï ou quelque chose comme ça + c'était une une pays XXX + ça comme ça c'est quelque chose qu'on + qu'on aime bien[5]

13 Dans l'exemple (1), l'apprenante, pour se faire comprendre, recourt à une stratégie qui selon nous peut admettre plusieurs interprétations et qu'il n'est pas possible de traiter à partir de la simple transcription orthographique standard (ci-après TOS) du corpus oral.

(2)
[INT] voilà la première image + vous pouvez la décrire et la commenter
[AP] est très belle premièrement {rire} nous pouvons voir la plage + il y a la mer + nous avons le ciel qui est très très clair + avec peu de nuages et qui est bleu + nous pouvons voir hm l'arbre + je ne sais pas comment on le dit en français mais c'est l'arbre qui est toujours à la plage {rire} euh nous pouvons voir un perroquet + hm je pense que c'est à Ibiza peut-être + ce n'est pas à Chypre cette image + ah + et + nous

pouvons voir une fille qui qui prend son dodo et + il y a des gens je pense qui nagent dans la mer mais généralement c'est une plage qui est assez vide de gens + c'est l'espace idéal pour les vacances je pense

14 En (2), nous pouvons remarquer que l'apprenante recourt systématiquement à la même tournure pour introduire les éléments de sa description. Son discours est tout à fait cohérent et suit indéniablement une progression, mais il est de force cohésive lâche (Cornish, 2003), ce qui rend difficile son évaluation à partir des seuls critères de l'échelle proposée par le CECRL.

(3)
[INT] bien voilà une première image + décrivez-la et dites ce que vous en pensez
[AP] première chose vacances + dans une plage dorée dans une mare dans un hm + très joli et hm + les palmes + où une fille qui qui se repose dessous de soleil et c'est la tranquillité totale {rire} + vacances + et dans une pays exotique bien sûr je sais pas comme euh ça pourrait être Mexique + Philippines ou quoi + le soleil + l'été
[INT] la deuxième
[AP] donc la musique + et + je sais pas si c'est la musique classique ou bien
[INT] décrivez-la aussi
[AP] oui et hm une + des des personnes qui + jouent jouent au vi- vo- + au violon c'est quoi + oui + et euh + oui un une personne de nationalité africaine + et une personne de nationalité européenne je crois + je sais pas + et c'est tout

15 Enfin, dans l'exemple (3), nous pouvons remarquer que l'apprenante interagit directement avec l'interviewer. Encore une fois, avec une simple TOS du corpus, il est difficile d'interpréter clairement l'intention de l'apprenante. Toutefois, il reste tout à fait envisageable d'interpréter ses questions comme un souci de pertinence par l'apprenante, qui cherche alors un feedback quant à ses hypothèses et à son développement. Cela pose la question de l'affectivité selon le cadre expérimental et son influence sur les données, sur laquelle nous revenons plus loin.

3.2. Quelques échantillons du corpus écrit

16 Proposons maintenant quelques exemples représentatifs du second corpus[6] en situation de production écrite.

(4)
En 1950, partir pour les vacances n'était pas encore quelque chose de quotidien, mais plutôt un luxe. C'est pourquoi, quand M. et Mme Leconte sont partis, toute la famille et les voisins étaient là pour dire au revoir. (Texte narratif)

17 En (4), nous notons l'emploi de l'expression introductrice du cadre du discours *En 1950* qui contextualise le propos. En outre, les relations fonctionnelles sont très clairement exprimées à l'aide des marques de cohésion.

(5)
Voici la gentille famille Martineau. [...] À gauche Céline ; six ans, la tête toujours un peu dans les nuages, [...]
Ensuite il y a Loïc, [...]
Puis nous avons Julie ; épouse de Loïc [...]
Finalement, à droite, le petit Julien. [...]. (Texte descriptif)

18 Dans l'exemple (5), les marques configurationnelles spécifiques de l'écrit laissent percevoir une organisation structurelle très nette, accentuée par les connecteurs temporels.

(6)
Henri Martineau aime bien tous les sports, surtout la navigation et la planche à

> voile. Malheureusement, il n'a pas beaucoup de motivation académique — un fait qui s'inquiète beaucoup ses parents.
> Emma Martineau a huit ans, et bien qu'elle soit jeune, est évidemment très intelligente. Elle ne regarde jamais la télévision, préferant la lecture. (Texte descriptif)

19 Enfin, dans le cas de (6), les reprises anaphoriques et le développement du discours à l'aide de tournures concessives sont là encore des phénomènes que nous n'avions pas (ou très peu) rencontrés dans les productions du corpus oral.

4. Interprétation des résultats et questionnements

20 Force est de constater que d'un corpus à l'autre, les attitudes adoptées par les apprenants varient de manière notable. Rappelons que les niveaux des apprenants ayant effectué le test de placement est avancé[7], tandis que les apprenants interrogés pour la constitution du corpus Emo-FLE sont de niveaux intermédiaire et avancé[8]. Tout en prenant en considération ce facteur, il reste selon nous peu probable que des apprenants de niveaux sensiblement similaires aient un écart aussi important entre l'oral et l'écrit concernant la maitrise des capacités discursives. Rappelons que le CECRL n'admet pas de différence entre les deux types de production. Bien qu'aucun de ces énoncés ne soit incohérent, si nous avions à infirmer ou à confirmer notre hypothèse sur le décalage potentiel entre les descriptions du CECRL et les compétences réelles d'un apprenant certifié à partir de ces données, le corpus Emo-FLE tendrait à nous orienter vers l'affirmative tandis que le corpus écrit semble plus proche de l'échelle prescrite par le CECRL.

21 En effet, dans les exemples (1), (2) et (3), que nos apprenants soient de niveau B1+, B2 ou C1, au niveau lexical, les articulateurs et connecteurs utilisés sont très limités. Du côté syntaxique, on peut noter que la plupart des phrases ne sont pas complètes, que les structures organisationnelles sont relativement peu complexes, voire absentes, et que l'ensemble des productions est assez court. Ainsi, quel que soit le niveau réel de nos apprenants (B1+/B2/C1), leurs compétences ne paraissent pas fidèles à la progression décrite par le CECRL. À l'inverse, dans les exemples (4), (5) et (6), on remarque des productions significativement plus longues (nous ne fournissons que de courts extraits ici par souci de synthèse) — y compris pour une tâche similaire telle que la description — et des moyens lexicaux et syntaxiques bien plus variés.

22 Comment trouver réponse à notre hypothèse avec des résultats si différents ? Les écarts de compétence constatés seraient-ils dus au canal de production ? En réalité, c'est là que nous revenons à l'objectif premier de notre article. Pour trouver des éléments de réponse, il est indispensable de s'interroger sur les contextes d'élaboration et de constitution de nos deux corpus. En effet, comme mentionné dans notre introduction, chaque corpus est constitué dans un but précis qui oriente les données obtenues (Bilger, 2008). Ainsi, un corpus neutre et objectif pour une recherche donnée peut tout à fait ne plus l'être pour une autre. Dans le cas de ces deux corpus, nul doute que le cadre expérimental est d'un enjeu majeur.

23 Revenons sur nos exemples (1), (2) et (3). Dans le cadre expérimental du corpus Emo-FLE, l'apprenant se retrouve face à un interviewer. La consigne d'origine est supposée permettre à l'apprenant de développer une production supposée monologale descriptive. Seulement, il est impossible pour la majorité des apprenants d'ignorer la

présence de l'interviewer. En effet, dans le cas de l'exemple (3), on remarque que l'apprenant pose des questions qui paraissent être directement adressées à son interlocuteur ou reste parfois dans l'attente d'une réaction de sa part, basculant ainsi sur une modalité dialogale interactionnelle, et ce même si l'interactant ne réagit pas (Rançon, 2016). N'oublions pas que dans la plupart des situations d'exercices oraux comme celui-ci, l'interviewer, et peut-être d'autant plus s'il est passif, peut se voir endossé l'étiquette d'évaluateur par l'apprenant. Ainsi, dans une telle situation, certains apprenants, dépendamment de leur degré d'affectivité, peuvent être en permanence en quête d'indices sur la qualité de leurs productions dans les yeux ou dans les expressions de ce dernier. Pour peu que celui-ci acquiesce machinalement et hoche la tête, les apprenants penseront alors potentiellement que leur discours est parfaitement clair et structuré et poursuivront sans souci de clarté. À l'inverse, si celui-ci reste de marbre, il se peut que l'apprenant soit pris d'incertitudes en pensant qu'il ne s'exprime pas clairement ou qu'il n'est pas pertinent et de ce fait passer à un nouveau topique sans raison apparente dans la transcription (Christoforou & Kakoyianni-Doa, 2014 ; Bogaards, 2007).

24 Au-delà de cette quête d'indices susceptible d'avoir un impact sur la production des apprenants, il est impossible de nier les éléments multimodaux et paraverbaux de la communication orale face à un interviewer. Ainsi, à l'inverse d'un corpus écrit où la plupart des informations quant aux stratégies employées sont observables, la seule TOS d'un corpus oral (en supposant donc que celui-ci n'ait pas été aligné sur vidéo à l'aide d'un logiciel) ne nous permet pas de repérer tous les moyens[9] auxquels l'apprenant recourt pour éclairer son propos. Rançon (2016) revient par exemple sur l'importance de prendre en compte ces aspects multimodaux s'il convient avec ce même corpus d'observer comment les apprenants manifestent leurs émotions. À l'aide de la vidéo, les gestes et expressions des apprenants sont décomptés et décryptés et les résultats laissent ressortir le fait qu'une grande partie du message est manifesté de manière paraverbale. Reprenons notre exemple (1) et intéressons-nous au passage suivant :

(1)
c'était une une pays XXX + ça comme ça c'est quelque chose qu'on + qu'on aime bien

25 Cette désignation (passage souligné) est extrêmement difficile à interpréter sans la vidéo. Il convient d'être très prudent quant à son analyse. En effet, il se peut que ce soit une reprise anaphorique, auquel cas la stratégie de l'apprenant pour être cohérent peut être identifiée comme l'un des moyens linguistiques recommandés par le CECRL. On peut ainsi l'interpréter de cette manière :

(1')
L'été avec le soleil, un endroit calme, Hawaï : c'est quelque chose qu'on aime bien.

26 Une autre interprétation fortement probable serait que l'apprenant est en train d'effectuer un geste de pointage vers l'image en référant ainsi à l'objet qu'il décrit tout en parlant. Dans ce cas, la stratégie utilisée par l'apprenant est d'ordre paraverbal, ce que le CECRL ne mentionne pas :

(1'')
Cette image que j'ai sous les yeux représente parfaitement quelque chose qu'on aime bien.

27 Dans les deux cas, l'analyse et l'interprétation des stratégies employées par l'apprenant pour être cohérent ne sont pas les mêmes et posent ainsi des questions allant au-delà du simple cadre expérimental et englobant le traitement des données. En effet, ne pas pouvoir interpréter avec certitude ce passage nous interpelle sur les choix de

transcription à effectuer pour évaluer les compétences des apprenants. La TOS seule du corpus peut-elle suffire ou faut-il aligner cette transcription avec l'audio ou la vidéo[10] ? En effet, l'échelle préconisée par le CECRL semble être parfaitement observable à partir de la TOS uniquement. Seulement, ne risque-t-on pas de passer à côté d'autres stratégies — non perceptibles via une TOS — utilisées par les apprenants ? Bien entendu, cela dépend de l'objet de l'étude et cette question rejoint les nombreuses réflexions déjà existantes en la matière (Bilger, 2008). En revanche, l'analyse de ces résultats ne laisse aucun doute quant au fait que dans cette situation spécifique, les apprenants ont recouru à d'autres stratégies que celles décrites par le CECRL.

28 Nous rejoignons les observations de Blanche-Benveniste (1997) et Cappeau (2008) sur le fait qu'il existe une syntaxe de l'oral et n'entendons pas ici dire que l'oral ne serait pas structuré. Seulement, dans le cas du corpus Emo-FLE, les apprenants se retrouvent en situation de production spontanée. Ils ne se voient pas offert de temps de préparation. Si nous regardons à présent le cas du test de placement, même si le temps est limité, les apprenants peuvent réfléchir à l'avance sur ce qu'ils vont écrire. Ils ont le temps de construire, de déconstruire, de reconstruire, voire de relire leurs énoncés pour s'assurer que leur propos est cohérent et organisé. L'écrit offre en outre des marques configurationnelles directement visibles dans les productions qui peuvent parfois aider à la hiérarchisation et à la cohérence du texte (alinéas, organisateurs métadiscursifs, etc.) (Charolles, 1994). Ainsi, il peut sembler normal que les productions de ce corpus nous paraissent plus structurées que celles du corpus Emo-FLE. La remarque que l'on pourrait alors émettre serait que si l'oral et l'écrit possèdent des systèmes syntaxiques différents, il pourrait alors être bénéfique de préciser les mécanismes de chacun pour la construction de la cohésion et de la cohérence, puisque les stratégies employées pour les construire sont d'autant plus susceptibles de différer d'un canal à l'autre.

29 Ajoutons également qu'à l'écrit, l'apprenant n'est pas intimidé ou influencé par la présence d'un interviewer (Bogaards, 2007). De ce fait, étant seul évaluateur de son propos, il est fort probable qu'il soit plus alerte quant à la clarté de celui-ci en ayant le souci d'être compris par le futur lecteur. Il peut alors se réfugier dans des tournures syntaxiques sécurisantes qu'il maitrise déjà à la manière des *lexical teddy bears* chez les apprenants de niveau avancé (Hasselgren, 1994).

30 Enfin, ces deux contextes auront sans doute été perçus différemment par chaque groupe d'apprenants. Ainsi, le test de placement, qui consiste très exactement en une évaluation, aura peut-être un statut plus formel pour les apprenants que le cadre expérimental proposé pour le corpus Emo-FLE. L'utilisation plus variée des connecteurs peut alors résider dans le fait que le contexte est un contexte évaluatif formel, auquel les apprenants ont certainement déjà été confrontés, et pour lequel les enseignements institutionnels et les outils pédagogiques recommandent cette variété lexicale et syntaxique. À l'inverse, nous ne disposons que de peu d'informations sur l'aspect formel et sur les enjeux sous-jacents aux productions des apprenants pour le corpus Emo-FLE, mais il est possible qu'une formalité moindre ait influé sur un niveau de langue moins soutenu, comportant des marques de cohésion lexicalement moins variées et un usage global moins formel. En l'absence de plus amples informations, il ne s'agit là que de simples hypothèses, mais nous pourrions tout à fait imaginer obtenir des indices plus précis quant à cette perception de la situation à l'aide d'un questionnaire d'enquête sociolinguistique par exemple. Quelques questions pourraient ainsi porter sur le degré de stress ressenti ou sur la perception de l'attitude de

l'interviewer, ce qui permettrait d'avoir une idée plus précise du vécu de la situation par les apprenants.

5. Conclusion et perspectives

31 Au vu de l'analyse des extraits sélectionnés, il convient de dire que les résultats ne permettent pas de confirmer ou d'infirmer l'hypothèse émise en introduction. L'expérience que nous venons de mener témoigne des limites de la réutilisation de corpus déjà existants. En effet, les corpus qui nous ont été prêtés pour la réalisation de cette expérience sont parfaitement viables scientifiquement au regard des objectifs de recherche initiaux pour lesquels ils ont été constitués. Le but de cet article n'est certainement pas de remettre cela en question, mais plutôt de sensibiliser quant au fait que le biais méthodologique vient souvent du chercheur lui-même. Parfois aveuglé par la praticité que représente un corpus déjà existant en termes de gain de temps, il peut rapidement oublier de vérifier si celui-ci présente bien toutes les informations et métadonnées nécessaires à sa recherche. Sur ce plan, il n'est pas rare que les corpus oraux constitués par des équipes de recherche ne soient que partiellement disponibles. Cappeau (2008, p. 21) souligne que tantôt les enregistrements sont disponibles, tantôt les transcriptions, mais rares sont les cas où les deux sont accessibles simultanément. Dans les quelques cas où les deux sont disponibles, il convient de rester tout aussi prudent. Nous avons vu par exemple qu'un enregistrement audio ne donne pas d'informations sur la configuration spatiale entre les interactants, ni sur leur mimogestualité. Or, tirer des conclusions à partir des seuls éléments observables, sans les remettre en question et sans prendre en considération les paramètres extérieurs ayant potentiellement une influence, est un raccourci dangereux sur le plan épistémologique.

32 De cette manière, nous espérons que cet article a permis de montrer à quel point il convient d'être prudent lors de l'élaboration ou lors de la réutilisation de corpus. Gardons à l'esprit que le corpus, en tant qu'outil de recherche, ne trouve sa viabilité scientifique que si le chercheur fait preuve de rigueur (Cori & David, 2008). Ne pas avorter les chances que les locuteurs recourent à des stratégies représentatives de leurs usages sans pour autant tomber dans une démarche circulaire devient alors le juste équilibre que tout chercheur s'engage à respecter. Pour cela, soumettre le protocole à de nombreux prétests et remettre systématiquement en question les données obtenues devraient permettre de pallier les risques mentionnés plus haut.

BIBLIOGRAPHIE

ABOUDA, Lofti & BAUDE, Olivier. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas ESLO. Dans C. Duteil-Mougel & B. Foulquié (dir.), *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation* (p. 155-162). Albi, France : Texto!

BILGER, Mireille. (2008). Avant-propos. Dans M. Bilger (dir.), *Données orales. Les enjeux de la transcription*, *Cahiers de l'université de Perpignan*, *37*, 8-11.

BILGER, Mireille & CAPPEAU, Paul. (2008). Conclusions et perspectives. Dans M. Bilger (dir.), *Données orales. Les enjeux de la transcription*, *Cahiers de l'université de Perpignan*, *37*, 289-293.

BLANCHE-BENVENISTE, Claire (dir.). (1997). *Approches de la langue parlée en français*. Paris : Ophrys.

BOGAARDS, Paul (dir.). (2007). *Aptitude et affectivité dans l'apprentissage des langues étrangères*. Paris : Hatier / Didier.

CAPPEAU, Paul. (2008). Corpus de langue parlée. État des lieux en France. Dans M. Bilger (dir.), *Données orales. Les enjeux de la transcription*, *Cahiers de l'université de Perpignan*, *37*, 18-23.

CARLSEN, Cecilie Hamnes. (2012). Proficiency Level—a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics*, *33*(2), 161-183.

CARTER-THOMAS, Shirley. (1999). La stratégie thématique : son importance dans l'analyse textuelle. Dans C. Clairis (dir.) *5e Journée de la formation doctorale de linguistique générale et appliquée* (p. 49-64). Paris : Université René Descartes – Paris 5.

CHAROLLES, Michel. (1978). Introduction aux problèmes de la cohérence des textes. *Langue française*, *38*, 7-41.

CHAROLLES, Michel. (1994). Cohésion, cohérence et pertinence du discours. *Travaux de linguistique*, *29*, 125-151.

CHRISTOFOROU, Nathalie & KAKOYIANNI-DOA, Fryni. (2014). Blocages et stratégies en expression orale : le cas des Chypriotes hellénophones FLE. Dans F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer & S. Prévost (dir.), *SHS Web of Conferences : 4e Congrès mondial de linguistique française*, (vol. 8, p. 915-926). Les Ulis, France : EDP Sciences. <https://doi.org/10.1051/shsconf/20140801367>.

CONSEIL DE L'EUROPE. (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris : Éditions Didier.

CORI, Marcel & DAVID, Sophie. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages*, *171*, 111-129.

CORNISH, Francis. (2003). Types de relations de discours entre énoncés : interactions avec l'anaphore transphrastique. *Cahier du CRISCO*, *12*, 69-84.

DANEŠ, Frantisek. (1974). Functional Sentence Perspective and the Organization of the Text. Dans F. Daneš (dir.), *Papers on Functional Sentence Perspective* (p. 106-128). Berlin : De Gruyter.

DELIC. (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé* , *18*, 11-42.

GADET, Françoise. (2008). Les corpus oraux et la diversité des productions langagières. *Verbum*, *XXX*(4), 261-273.

HASSELGREN, Angela. (1994). Lexical Teddy Bears and Advanced Learners: A Study into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237-258.

RABATEL, Alain. (2013). L'engagement du chercheur, entre « éthique d'objectivité » et « éthique de subjectivité ». *Argumentation et analyse du discours*, *11*. <https://doi.org/10.4000/aad.1526>.

RANÇON, Julie. (2016). La multimodalité lors de l'expression d'émotions chez les Chypriotes allophones. Dans R. Nita & F. Valetopoulos (dir.), *L'expression des sentiments : de l'analyse linguistique à l'application* (p. 339-358). Rennes, France : Presses universitaires de Rennes.

REINHART, Tanya. (1980). Conditions for Text Coherence. *Poetics Today*, *1*(4), 161-180.

SALLES, Mathilde. (2006). Cohésion-cohérence : accords et désaccords. *Corela*, *HS-5*. <https://doi.org/10.4000/corela.1426>.

SINCLAIR, John (dir.). (1991). *Corpus, Concordance, Collocation*. Oxford, Angleterre : Oxford University Press.

VIGIER, Denis. (2012). Linguistique textuelle et enseignement du FLES. *Le français dans le monde. Recherches et applications*, *51*, 34-49. Disponible en ligne sur <https://halshs.archives-ouvertes.fr/halshs-00801532v2/document> (consulté le 3 décembre 2018).

NOTES

1. En France, l'absence de très grands corpus de référence et le fonctionnement par petites équipes sur des corpus non disponibles limitent les réutilisations massives. En revanche, les corpus disponibles voulus de référence ne rendent leurs données que partiellement disponibles (Abouda & Baude, 2006 ; Cappeau, 2008).

2. Au singulier dans le CECRL.

3. Ce corpus a été initié dans le cadre du projet international « Les sentiments à travers les corpus d'apprenants » en faisant l'objet d'un financement ACI et est toujours en cours de constitution.

4. Également en cours de constitution.

5. La transcription a été réalisée en s'appuyant sur la convention du groupe de recherche DELIC (2004). Néanmoins les éléments soulignés ne sont pas liés à cette convention, il s'agit de nos propres soulignements pour mettre en exergue les passages qui feront l'objet d'analyses.

6. Pour la transcription de ces extraits écrits, les signes de ponctuation, les marques configurationnelles (tels que les retours à la ligne) et les écarts graphiques sont fidèles à la production écrite papier des apprenants. En revanche, comme pour les extraits du corpus oral, les soulignements ont été ajoutés par nous-même autour des éléments sur lesquels les analyses porteront.

7. En cours d'acquisition des niveaux C1 ou C2.

8. D'après les corps enseignants de leurs établissements respectifs. Or, depuis les travaux de Carlsen (2012) on sait que le niveau de compétence peut être jugé totalement différemment d'un établissement à un autre, rendant cette variable peu fiable sans établir une évaluation à priori ou à posteriori.

9. Comme le recours à la multimodalité pour se faire comprendre via des moyens non linguistiques par exemple.

10. Nous pouvons par exemple penser à l'utilisation d'un logiciel tel que PRAAT pour un alignement audio ou tel que ELAN s'il est nécessaire d'aligner la transcription sur vidéo.

RÉSUMÉS

La linguistique sur corpus n'en est plus à ses débuts et pourtant pour un jeune chercheur, les pièges à éviter lors de la constitution ou de la réutilisation de ces corpus sont nombreux. Les corpus d'apprenants et leur exploitation, quel qu'en soit l'objectif, ne dérogent pas à ce constat. Dans cet article, nous essayerons, en tant que jeune chercheur, de nous interroger sur l'impact de la méthodologie de constitution du corpus et sur les analyses qui en découlent. Ce faisant, nous entendons sensibiliser tout chercheur à ces enjeux en invitant chacun à mieux les considérer lors de la mise en place de ses propres cadres expérimentaux.

Corpora analysis in linguistic research has already proven its efficiency. However, the elaboration of new corpora or the reuse of existing corpora can be tricky for a young researcher. Learner corpora and their analysis are no strangers to all the issues that can occur. In this article, we—as a young researcher—will consider how the methodology in elaborating new corpora can impact the data and their analysis. In doing so, we aim at making researchers aware of some of the potential biasses when elaborating their own experimental framework.

INDEX

Keywords : corpus linguistics, corpus constitution, learner corpora, coherence-cohesion
Mots-clés : linguistique sur corpus, élaboration de corpus, corpus d'apprenants, cohérence-cohésion

AUTEUR

ANTONIN BRUNET

FoReLLIS (EA 3816), Université de Poitiers